Mining process factor causality links with multi-relational associations

Mickael Wajnberg, Petko Valtchev, Mario Lezoche, Hervé Panetto, Alexandre Blondin Masse

Mining Process Factor Causality Links with Multi-relational Associations

Mickael Wajnberg
Dept. d'informatique, UQAM, Montréal, Canada
Université de Lorraine, CNRS, CRAN, Nancy, France
wajnberg.mickael@courrier.uqam.ca

Petko Valtchev
Dept. d'informatique, UQAM, Montréal, Canada
valtchev.petko@uqam.ca

Mario Lezoche, Hervé Panetto
Université de Lorraine, CNRS, CRAN, Nancy, France
{mario.lezoche,hervé.panetto}@univ-lorraine.fr

Alexandre Blondin Massé
Dept. d'informatique, UQAM, Montréal, Canada
blondin_masse.alexandre@uqam.ca

ABSTRACT

To make knowledge-supported decisions, industrial actors often need to examine available data for suggestive patterns. As industrial data are typically unlabeled and involve multiple object types, unsupervised multi-relational (MR) data mining methods are particularly suitable for the task. Current MR association miners merely produce singleton-conclusions rules hence might miss multi-way dependencies. Our novel MR miner builds upon a relational extension of concept analysis to extract general associations. While successfully dealing with circularity in data, it avoids producing cyclic rules by limiting the description depth of relational concepts. Our rules' relevance was validated by an application to aluminum die casting.

KEYWORDS

Relational datasets, association rules, concept analysis, industrial processes

ACM Reference Format:
Mickael Wajnberg, Petko Valtchev, Mario Lezoche, Hervé Panetto, and Alexandre Blondin Massé. 2019. Mining Process Factor Causality Links with Multi-relational Associations. In *Proceedings of Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3360901.3364446

1 INTRODUCTION

Understanding the domain behind the data is a key to business growth and competitiveness. Knowledge discovery from data (KDD) helps addresses that concern by distilling trends and patterns that are intelligible to human experts[5]. In industry, data objects are typically unlabeled and often comprise both proper features and object-to-object links. Such datasets fit the unsupervised multi-relational data mining (MRDM) mode [4], i.e. clustering and association discovery. However, existing MRDM association miners [3, 6, 8] restrict their output format to singleton-premise rules, hence they fail to capture more subtle associations.

Formal concept analysis (FCA) [7] has been proven as a versatile framework for KDD [12] in many practical applications [2]. It extracts knowledge as a compact set of association rules [10]. Relational concept analysis (RCA) [11] is MRDM extension of FCA. However, straightforwardly-defined relational association rules may easily contain circular references or references from conclusion to premise, thus preventing a meaningful interpretation. In this article, we illustrate an untangling method to avoid definition cycles that trims concept descriptions in RCA.

As a validation, we applied our method on industrial manufacturing data. The goal was to assist a domain expert who examines the production process for potential optimizations. At a first step, the expert searches for causality links between process factors and product anomalies. Our method supports the task by providing associations between machine state descriptors and product (qualitative) metrics. In a concrete experiment, a fair number of the discovered rules were deemed unexpected yet relevant by the experts involved.

In the remainder of the paper, section 2 motivates our study. Then, section 3 provides background while sections 4 and 5 describe our association mining approach and the experimental study, respectively. Section 6 discusses our results and section 7 concludes.

2 MOTIVATION

In many industrial contexts, the root challenge is finding the best trade-off between product quality, working time and manufacturing costs. In looking for a solution, it is crucial to reflect the risk factor [13], which can be assessed by constantly monitoring the machining process, e.g. with sensors and data analysis. Our case study covers an aluminum die casting process whose output is door/window handles and frames. In the partner workshop, process monitoring consists in regularly controlling product metrics. For instance, when a product is discarded for non-compliance to quality standards, the operator halts the machine, fixes the observed problem, and restarts the production whereby the product is melted again and reprocessed. As shown below, an in-depth analysis of production logs reveals regularities in the form of associations between variations in product measures, machine state and production issues. While few associations reflect true causality, many others still help understanding and, potentially, avoid machine failures and reduce costs.

K-CAP '19, November 19–21, 2019, Marina Del Rey, CA, USA

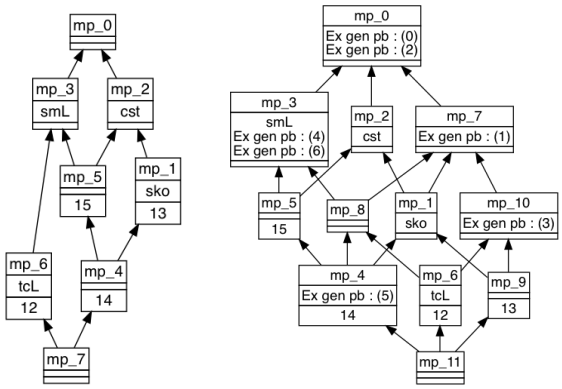
ACM ISBN 978-1-4503-7008-0/19/11…$15.00
https://doi.org/10.1145/3360901.3364446

K_1	*sko*	*cst*	*smL*	*tcL*
12			×	×
13	×	×		
14	×	×	×	
15		×	×	

g	P_0	P_1	P_2
12	×		×
13	×		
14		×	×
15			×

K_2	*t*5	*stp*	*qlt*	*mld*	*cost*
P_0	×	×		×	×
P_1	×		×		×
P_2		×	×	×	

Table 1: Relational Context Family of the machine part dataset.



(a) Without scaling. **(b) After one step of scaling.**

Figure 1: Machined part lattices (K_1).

3 BACKGROUND ON FCA AND RCA

Formal concept analysis [7] is an algebraic approach for eliciting the conceptual structure of a dataset. Input data format is a triple $K = (O, A, I)$ called a (formal) *context*, e.g. K_1 in Table 1. Here O is a set of objects (machined parts 12 to 15), A a set of attributes (*sko* for inadequate thickness, *smL* for thickness, *tcL* for pressure below threshold, *cst* for need recasting) and $I \subseteq O \times A$ an incidence relation listing valid pairs (o, a) (object o has the attribute a). FCA reveals all pairs of sets $(X, Y) \in \wp(O) \times \wp(A)$ strongly correlated, meaning that all objects having the attributes in Y are in X and *vice-versa*. Such pair is a (formal) *concepts* with an *extent* X and *intent* Y. For instance, $(\{13, 14\}, \{sko, cst\})$ is a concept, but $(\{14, 15\}, \{sko, cst, smL\})$ is not. Concepts are partially ordered w.r.t. extent inclusion $(X_1, Y_1) \leq_K (X_2, Y_2)$ iff $X_1 \subseteq X_2$ whereby the underlying hierarchy is a complete lattice. Fig. 1a depicts the Hasse diagram of the lattice derived from Table 1. It uses reduced concept labeling: Extent-wise (resp. intent-wise), a concept "inherits" the objects located at any sub-concept (resp. super-concept). For instance, the concept mp_5 is $(\{14, 15\}, \{smL, cst\})$.

An association rule is a pair $(Y, Z) \in \wp(A) \times \wp(A)$ written $Y \rightarrow Z$ [1]. It embodies information about co-occurrences of Y and Z in objects from O. Two classical evaluation metrics for associations are support (percentage of objects incident to $Y \cup Z$) and confidence (percentage of objects with Z among those with Y). Here, we focus on 100% confidence rules (a.k.a. implications). Furthermore, our rules optimized in that they have the form $Y \rightarrow Z - Y$ where Z is a concept intent whereas Y is a minimal subset of Z with the same support (a.k.a. generator) [10].

K_2	...	$\forall\exists g^{-1} : mp_0$	$\forall\exists g^{-1} : mp_1$	$\forall\exists g^{-1} : mp_2$	$\forall\exists g^{-1} : mp_3$	$\forall\exists g^{-1} : mp_4$	$\forall\exists g^{-1} : mp_5$	$\forall\exists g^{-1} : mp_6$
P_0	...	×						
P_1	...	×	×	×	×	×	×	
P_2	...	×			×			

Table 2: Extended problem context.

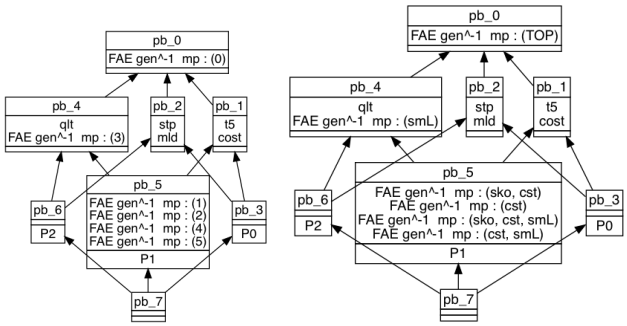
Relational concept analysis [11] assumes datasets are made of several contexts, one per type of object, and context-to-context relations, i.e. sets of object-level links. Within our production case, Table 1 depicts such a dataset (called *relational context family*): Here, K_2 describes production problems via attributes such as "fixing time less than 5 min" (*t*5), "machine stopped" (*stp*), "quality-related" (*qlt*), "mould defect" (*mld*), or "medium financial impact" (*cost*). A relation "generates" (g) links machine parts to observed problems. RCA uses *propositionalization* [9] to turn links into dedicated sort of attributes called *relational*. In doing that, it uses *scaling operators* akin to value restrictions in description logics and concepts from the relation's range context. The underlying format is $q\ r : (c)$ where $q \in \{\exists, \forall, \forall\exists, \ldots\}$ (with intuitive semantics [11]), r is the relation and c a concept. Incidence between objects of the domain context and its new relational attributes depends on the scaling operator used. Table 2 illustrates the extension of K_2 after scaling with $\forall\exists$ upon "is-generated-by" (denoted g^{-1}). Each concept from K_1 yielded a relational attribute incident to specific objects in K_2. For instance, problem *P*2 is related to parts 12, 14, and 15, whose common concepts are mp_0 and mp_3 (see Fig. 1a). Thus, two new attributes, $\forall\exists g^{-1} : (mp_0)$ and $\forall\exists g^{-1} : (mp_3)$, are assigned to *P*2.

After scaling relations with constructors, their respective domain contexts get extended. This knowingly leads to a set of *new* concepts popping up in the lattices of the updated contexts (see Figs. 1b and 2a). Since they represent additional abstractions, a scaling step will turn new concepts into yet newer attributes, and the whole cycle would go on. The overall iterative method of RCA provably reaches a fixpoint made of a set of inter-related lattices [11].

4 RCA-BASED KNOWLEDGE DISCOVERY

RCA is yet to be provided with a generic notion of association rule: The difficulty here lies in the references between relational concepts which might lead to circular dependencies. For instance, a rule extracted from pb_4 would be $qlt \rightarrow \forall\exists g^{-1} : mp_3, \forall\exists g^{-1} : mp_0$, when replacing a reference by the concept intent, mp_3 would become $smL, \exists g : pb_6, \exists g : pb_0, \exists g : mp_2, \exists g : mp_4$, the latter attribute would be replaced by pb_4 intent, establishing a circularity.

To prevent this, we modify slightly the original algorithm by exploiting the iterative nature of RCA which is cycle-free. At creation time, the intent of a content can only hold static attributes and relational ones referring to concepts created at previous iterations. Therefore, when replacing a reference to concept c in a relational attribute, we use only the "birth" intent. If any relational attributes exists in the birth intent c, then we recursively use the birth description of the referred concept, which has necessarily been created in a previous iteration. Finally, any relational intent can be described with only non relational attributes. Such expansion avoids circular dependencies, even if one may exist between full intents.



(a) Direct references. **(b) Disentangled references.**

Figure 2: Problem lattices after one scaling step.

On the same example, the obvious circularity is avoided if mp_3 is replaced by its intent at the end of its creation iteration (see Fig. 1a). The resulting intent of pb_4 is drawn in Fig. 2b.

5 EXPERIMENTAL STUDY

We now focus on an industrial case of aluminum die casting, whose data have been provided by Master Italy SRL, a company specialized in the manufacture of handles and frames for doors and windows.

Given the size of its activities, controlling, cleaning and managing fixes for the machine are costly operations. Thus, the manufacturer wishes to minimize these actions. To do so, instead of targetting the machines, the data focus on the products themselves. When a machined part fails to reach the required standards, it is put aside and eventually recast. Whenever the operator assesses the number of defective parts as being too high, the machine is stopped for a fix. Such policy allows the number of stops to be kept low. However, if the causality dependencies for problems could be identified, e.g. in terms of variations in parts properties, potential failure could be anticipated or completely avoided, along with the recasting.

To find such causality links, correlations between problems and part features are measured. Since the relation between machined parts and problems is many-to-many, a MRDM approach seems a natural choice, and since data are unlabeled, only descriptive approaches such as clustering or association mining are possible, hence our focus on RCA.

5.1 Dataset and experimental setting

Our experimental data cover one month of production, i.e. approximately 58.000 parts. In the present experiments, we focus on the 5.134 parts that relate to at least one problem. The raw data is as follows. A first table represents defective machined parts together with their static attributes, i.e. a total of 25 features tracked, among those, the production period, the mould, the product dimensions, the piston state at the end of each manufacturing step. A second table stores the 19 problems categories: Each is described by its nature (mechanical wear, machine calibration, etc.), the induced financial and time losses. A third table gathers the relational links (many-to-many) between the two previous tables: It indicates, given a machined part, all the related problem.

The two first tables have been scaled to form binary formal contexts: Categorical values are divided into all possible categories and numerical value are divided into five equal intervals. It resulted in 134 attributes for the machined part context and 26 for the problem one. Attributes are of the form $a_{i\_j\_k}$, where a is the target feature, i is the manufacturing process step, j is the interval (1 being the lowest value, 5 the highest). A final element k, optional depending on a (valued *ko*, *ok* or *check*), is a complementary description for the interval j. It indicates whether values in j should induce problems (*ko*) or not (*ok*), whereby *check* provides no systematic information.

The relation "generates" is scaled with $\exists$: If a concept c has $\exists gen : (\bar{c})$ in its intent, then every machined part in the extent of c generates at least one problem that has all the attributes in the intent of $\bar{c}$. The inverse relation "is_generated_by" is scaled with the $\forall\exists$ operator. If a concept c has $\forall\exists gen^{-1} : (\bar{c})$ then every problem of c is only related to machined parts having the attributes in the intent of $\bar{c}$. The operator $\forall\exists$ was chosen to provide a summary of the features shared by related parts (rather than multiple cases as with $\exists$).

Next, only association rules of sufficient support (threshold set to 20) were extracted. Moreover, further filters on rules eliminated those with no relational attribute. Finally, at this preliminary step, we only examined 100% confidence rules.

5.2 Analysis outcome

Among the 133.821 rules output by our RCA tool, we selected a small number that seemed to be the most valuable from an expert point of view. Table 3 shows some of the selected rules which are split into groups for which we provide expert interpretation.

Multiple problems related to the same machined parts. RCA found 3.950 rules reflecting the fact that machined parts may be related to multiple problems. This is a rather rare situation, hence it was interesting for the expert to see typical co-occurrences between problems and their characteristics, as well as those of the related machined parts. For instance, the first rule in Table 3 has a 7% support. It states that any machined part which reaches the lower bound ($min\_sm(< LimLow)$) of the lowest interval for the thickness (sm_1_ko) will have to be recast. Unsurprisingly, it will generate an alert as the thickness lays below the 14mm threshold (the problem will be recorded as such). Less trivially, the rule indicates that, invariably, a micro-stop will be triggered to deal with thickness issues, which was judged by the machine operator as being predictable.

Recasting conditions. Another point of interest were combinations of machined part attributes that could be an indicator for future recasting needs. As a first, and rather direct approach, we selected all rules comprising the *pdt_recast* attribute. Within the resulting set of 42.060 rules, we found a number of subsets of common concept. Lot #2 in Table 3 is an example of such subset whose support, 14%, speaks in favor of a recurrent phenomenon that deserves attention. While discussing the complete interpretation of the situation is beyond the scope here, it might be summarized as follows. A first remark is that very low temperature ($t_{2\_1\_ko}$) is a constant in the premise. It must be combined with either very high speed of piston at step one ($v_{1\_5\_ko}$) or piston course in the lowest values at step two ($c_{2\_1\_ko}$). The remainder of the premise brings the recast information. The conclusion invariably presents the piston course at step one in the highest interval ($c_{1\_5\_ko}$). Taking into account the specific combinations of attributes involved in the rules, both those present and some of the missing ones, led the expert team to the following hypothesis: After restarting, it takes

Lot #	Premise	Conclusion
1	$sm\_1\_ko, min\_sm(< LimLow)$	$pdt\_recast, \exists$ gen pbs : $(sm \leq 14)$, $\exists$ gen pbs : $(microstop - smLimLow, predictable)$
2.1	$c_{2\_1\_ko}, t_{2\_1\_ko}, pdt\_recast$	$c_{1\_5\_ko}, v_{1\_5\_ko}, \exists$ gen pbs : $(\forall \exists \text{ gen}^{-1}$ parts : $(pdt\_recast))$
2.2	$c_{2\_1\_ko}, t_{2\_1\_ko}, \exists$ gen pbs : $(\forall \exists \text{ gen}^{-1}$ parts : $(pdt\_recast))$	$c_{1\_5\_ko}, v_{1\_5\_ko}, pdt\_recast$
2.3	$v_{1\_5\_ko}, t_{2\_1\_ko}, pdt\_recast$	$c_{1\_5\_ko}, c_{2\_1\_ko}, \exists$ gen pbs : $(\forall \exists \text{ gen}^{-1}$ parts : $(pdt\_recast))$
2.4	$v_{1\_5\_ko}, t_{2\_1\_ko}, \exists$ gen pbs : $(\forall \exists \text{ gen}^{-1}$ parts : $(pdt\_recast))$	$c_{1\_5\_ko}, c_{2\_1\_ko}, pdt\_recast$
3	$\exists gen\ pbs : (rul\_component, down\_time\_1h+, phm)$	$pdt\_conform$
4	$v_{2\_4}, fc_{2\_ok}, \exists$ gen pbs : $(\forall \exists \text{ gen}^{-1}$ parts: $(10A026, c_{1\_5\_ko}, c_{2\_1\_ko}, v_{1\_5\_ko}, t_{2\_1\_ko}, c_{c\_2\_ok}, pm_1, sm_{2\_check}))$	$10A026, c_{1\_5\_ko}, c_{2\_1\_ko}, v_{1\_5\_ko}, t_{2\_1\_ko}, c_{c\_2\_ok}, pm_1, sm_{2\_check}$, $pdt\_conform$

Table 3: Sample rules from RCA output.

some time for the machine to reach the optimal temperature. Even if the metal is completely molten, the piston is not dilated enough, hence friction is low and speed high. This often results in extreme values for its course, hence the non conform parts that need recast.

Conform parts related to problems. Given that not all rules apply on products so defective that they need a recast, experts were also eager to examine the circumstances under which machined parts, albeit associated to problems, still remained conform to the norms. This is a valuable question, as answers may help identify indicators for upcoming issues. Indeed, some produced parts, while still conform, tend to have *some* of their feature values ever closer to the conformity limits. Identifying typical combinations of such values per problem category would be a significant advancement. We looked at rules with the $pdt\_conform$ attribute (53.008 rules). For instance, rule #3 in Table 3 indicates that if a machined part relates to a mechanical problem ($rul\_component$) which entails an hour or longer downtime ($down\_time\_1h+$) and to a preventive maintenance (phm), then the product would be necessarily conform. This is coherent because the machine stop is caused by a mechanical component failure and the last product manufactured is put in relation to the observed problem. This yields key insight into the state of the machine just before the problem showed up.

Problem description precision Finally, while looking for the conform parts, we discovered a small percentage of rules which, rather than bringing new insights, point out to deficiencies in the problem description. For instance, rule #4 in Table 3 (of support just below 1%) comprises a number of features which seem plausible plus a reference to a concept of problems. The latter, however, comprises a single relational attribute pointing to a super-concept of the one the rule is stemming from (carrying a strict subset of the static attributes). It is noteworthy that the absence of static attributes in that intent means at least two problems are in the extent. Now, our experts found it surprising that these problems, despite the highly similar profiles of the related parts (eight shared attributes) do not share a single static feature themselves. This is potential indication for missing such features in problem description (e.g. outside temperature that might influence the entire process).

6 DISCUSSION

RCA arguably succeeded in finding correlations of features between machined parts and problems. From its output, experts could detect rules discriminating problems that invariably force a recasting of the related parts from those which do not impede correct manufacturing. In a different vein, relational attributes naturally cluster machined parts that relate to the same category of problems, which highlights non relational attributes connected to a problem category. This narrowing of the search scope facilitates the task of detecting causality links between events for the experts.

Finally, while our rules yielded non trivial insights on machined parts and the way they relate to problems, the knowledge about problems gleaned from them was rather coarse-grained. Given the granularity of problem descriptions, this came as no surprise. Still, even such imperfect description was enough for RCA to extract hints for the industrial team, e.g. as to where new sensors, if installed, could have the biggest impact on analysis scope and depth.

7 CONCLUSION

We define relational association rules in a way that avoids improper references in rule parts without restricting their size. This unlocked the entire spectrum of FCA-based association mining mechanisms to arbitrary MR datasets. Validation study indicates that our method is capable of detecting non-trivial facts that are beyond the reach of competing approaches. Still, a larger evaluation effort will be necessary to solidify these preliminary findings, e.g. focusing on rules with less than 100% confidence.

REFERENCES

[1] R. Agrawal et al. Fast discovery of association rules. *Advances in knowledge discovery and data mining*, 12(1):307–328, 1996.
[2] M. d'Aquin and E. Motta. Extracting relevant questions to an rdf dataset using formal concept analysis. In *6th K-CAP conference*, pages 121–128, 2011.
[3] L. Dehaspe and H. Toivonen. Discovery of relational association rules. In *Relational data mining*, pages 189–212. Springer, 2001.
[4] S. Džeroski. Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter*, 5(1):1–16, 2003.
[5] U. Fayyad et al. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.
[6] L. Galárraga et al. Fast rule mining in ontological knowledge bases with AMIE+. *The VLDB Journal*, 24(6):707–730, 2015.
[7] B. Ganter and R. Wille. *Formal concept analysis: mathematical foundations*. 1999.
[8] B. Goethals and J. Van den Bussche. Relational association rules: getting warmer. In *Pattern Detection and Discovery*, pages 125–139. Springer, 2002.
[9] S. Kramer et al. Propositionalization approaches to relational data mining. In S. Džeroski and N. Lavrač, editors, *Relational Data Mining*, pages 262–291. 2001.
[10] M. Kryszkiewicz. Concise Representations of Association Rules. In *Pattern Detection and Discovery*, volume 2447, pages 92–109. Springer, 2002.
[11] M. Rouane-Hacene et al. Relational concept analysis: mining concept lattices from multi-relational data. *Annals of Mathematics and A. I.*, 67(1):81–108, 2013.
[12] P. Valtchev et al. Formal Concept Analysis for Knowledge Discovery and Data Mining: The New Challenges. In *Proc. of ICFCA 2004*, LNCS, pages 352–371.
[13] Z. Wang et al. Framework for modeling operational uncertainty to optimize offsite production scheduling of precast components. *Automation in Construction*, 86:69–80, 2018.